How much and how many?
Guidance on the extent of validation/verification studies

S L R Ellison

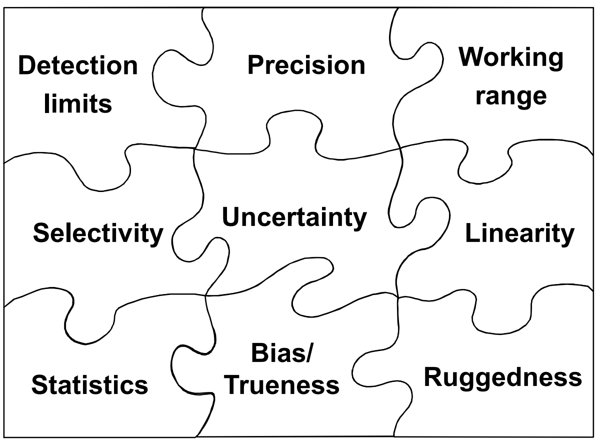
Science
for a safer world

Introduction

- Performance characteristics
  - How many need to be examined?
- Experiment size
  - How many samples and replicates are needed?
- Minimising the workload:
  - Multiple characteristics from single studies
  - Maximising information with efficient experiments

How many performance characteristics need to be examined?

A validation puzzle



Typical guidance on characteristics for study (ICH)

Performance Characteristics	Type of analytical procedure: IDENTIFICATION	TESTING FOR IMPURITIES quant	TESTING FOR IMPURITIES limit	ASSAY
Accuracy	-	+	-	+
Precision Repeatability	-	+	-	+
Interm.Precision	-	+ (1)	-	+ (1)
Specificity (2)	+	+	+	+
Detection Limit	-	- (3)	+	-
Quantitation Limit	-	+	-	-
Linearity	-	+	-	+
Range	-	+	-	+

ICH Q2(R1) (1994)

Typical guidance on characteristics for study (Eurachem)

Table 3 – Extent of validation work for four types of analytical applications. Example from the pharmaceutical sector [13]. 'x' signifies a performance characteristic which is normally validated.

Performance characteristic	Type of analytical application: Identification test	Quantitative test for impurity	Limit test for impurity	Quantification of main component
Selectivity	x	x	x	x
Limit of detection			x	
Limit of quantification		x		
Working range including linearity		x		x
Trueness (bias)		x		x
Precision (repeatability and intermediate precision)		x		x

NOTE The table is simplified and has been adapted to the structure and terminology used in this Guide.

Typical guidance on characteristics for study (IUPAC)*

Performance Characteristics	Previous validation		
	Full[1]	**Full[1]** New matrix	**Basic** (Literature)
Bias	✓	✓	✓
Repeatability	✓	✓	✓
Reproducibility	✓	✓	✓
Linearity	?	?	✓
Ruggedness	-	-	✓
Detection limit	Not mentioned – depends on use		

Note 1. "Full" validations includes collaborative study

** IUPAC Harmonised guidelines on single-laboratory validation*
Selected examples for quantitative analysis shown

Performance characteristic requirements

- Broadly similar across sectors
- Bias/trueness, precision and linearity always required for quantitative methods
- Detection capability usually examined
- Ruggedness requirements depend on sector
  - All agree that ruggedness can be useful in development
  - Some require ruggedness as part of a standardised validation suite

Experiment size

How many observations?

Selected guidance on experiment size

Performance Characteristics	Guidance document		
	ICH Q2	IUPAC SLV	Eurachem
Bias/Trueness	3 levels in triplicate	-	10 replicates**
Repeatability	3 levels in triplicate	-	6 – 15 replicates**
Reproducibility	-	-	6 – 15 in duplicate**
Linearity	5 levels	6 levels in duplicate	6-10 levels 2-3 times each
Detection limit	-		10 replicates
Ruggedness*	- †‡	- ‡	- ‡

'-' No numerical guidance given
* 'Robustness' in ICH guidance

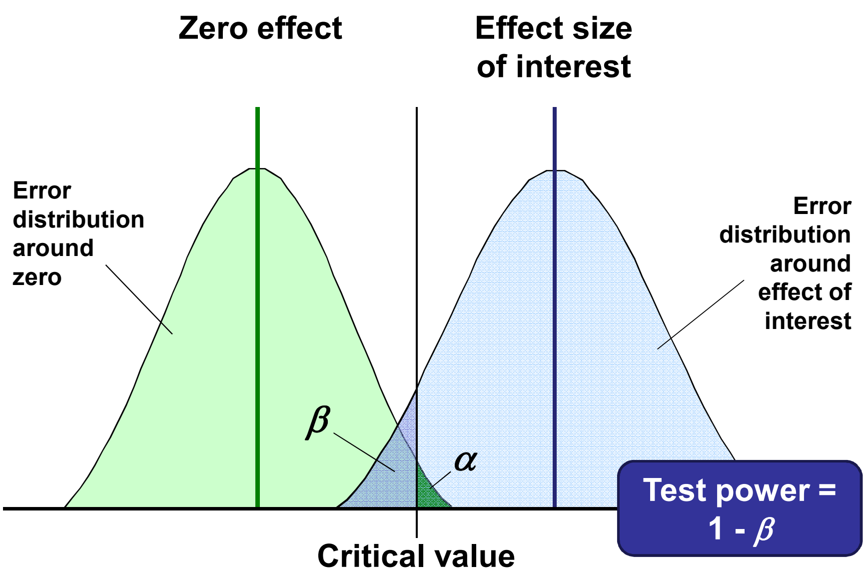
† Example conditions suggested
‡ Experimental designs suggested
** Per concentration/material studied

Test power for sample size calculation

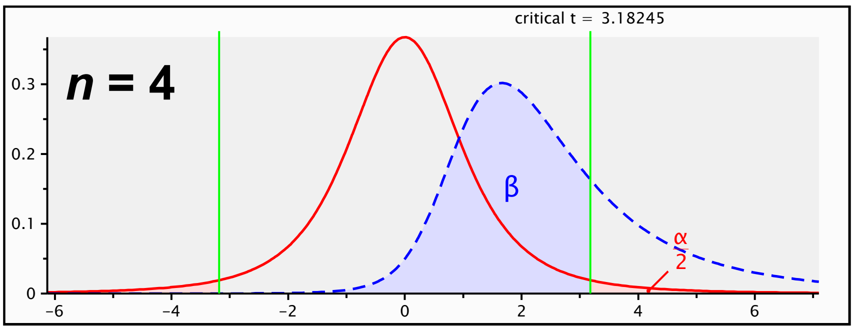
Some nomenclature

- Type I error: Incorrect rejection of the null hypothesis
  - Concluding there is an effect when there is none. A false positive.
- Type II error: Incorrect acceptance of the null hypothesis
  - Failing to find a real effect; a false negative.
- Power (of a test): The probability of correctly rejecting the null hypothesis when it is false.
  - Equal to 1 minus the Type II error probability.

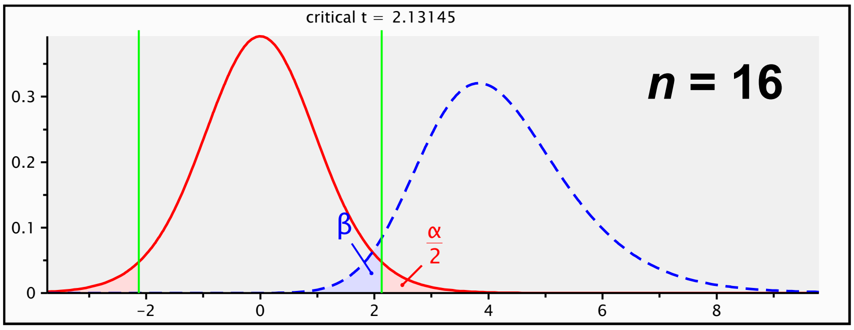
The concept of test power



Power for some *t* tests



Power for some *t* tests



It takes 16 observations to find a bias as small as 1 standard deviation (with 95% power)

Calculating test power: Required information

- A calculation of minimum sample size for a given test power requires:
  - a) The type of test (*t*-test, *F*-test *etc*.) and the details (One- or two-tailed? etc);
  - b) The size of the effect that is of interest;
  - c) The typical standard deviation *s*;
  - d) The required level of significance for the test (the Type I error probability α) and
  - e) The desired test power, usually expressed in terms of the probability β of a Type II error.
    - Typically 80% or 95%

Test power basis for bias experiments

Number of observations for 95% power at 95% confidence.

δ/s	0.5	0.6	0.7	0.8	0.9	1.0	1.5	2.0	2.5	3.0
n	55	39	29	23	19	16	9	6	5	4

NIST special publication 829: Use of NIST Standard Reference Materials for decisions on performance of analytical chemical methods and laboratories

δ = Size of bias to be detected
s **= Available precision**
n **= Number of observations required**

Power for precision experiments

- Can be calculated where a hypothesis test is intended
  - Chi-squared test for significantly exceeding required precision
  - *F* test for different precision in two groups
- Typical experiments are not very powerful for detecting excess dispersion
  - Detecting 40% excess dispersion* requires 7 replicates at 80% power and 18 at 95% power

* If the required precision is σ, true precision of 1.41σ will give a positive chi-squared test result 80% of the time with 7 replicates

Caveats

- Power calculations rely on assumptions
  - Likely effect size
  - Available precision
  - Distribution under the ‘alternate’ hypothesis
- These assumptions may be quite poor
- Power analysis is very useful for comparing designs under similar assumptions
  - ... but don’t over-interpret

Future directions

- Draft IUPAC guidance:
  Experiments for Single Laboratory Validation Of Methods of Analysis: Harmonized Guidelines
- Sets 3 levels of ‘stringency’
  - Verification, validation, stringent validation
- Provides ‘model experiments’
- Permits any other experiment that gives the same test power
- Gives guidance on number of materials, replication level, size of experiment and ‘stringency’ of validation

Draft IUPAC guidance – experiment size

Table 1: Minimum replication requirements			
Performance Characteristic	**Verification**	**Standard validation**	**Stringent validation**
Applicability			
Selectivity	See Table 2 note 3	4 replicates each on control and interferent-spiked material[Note 1]	7 replicates each on control and interferent-spiked material[Note 1]
Calibration linearity	4 levels in duplicate	6 levels in duplicate	Either 10 levels in duplicate *or* 5 levels in triplicate
Trueness and/or Recovery	6	10	16
Precision:			
Repeatability	3	7	18
Run-to-run (within-laboratory reproducibility) using simple replication	3	7	18
Run-to-run (within-laboratory reproducibility) using nested design	3 groups of 2	5 groups of 2	12 groups of 2

Look out for IUPAC consultation

Part 2: Getting more for less

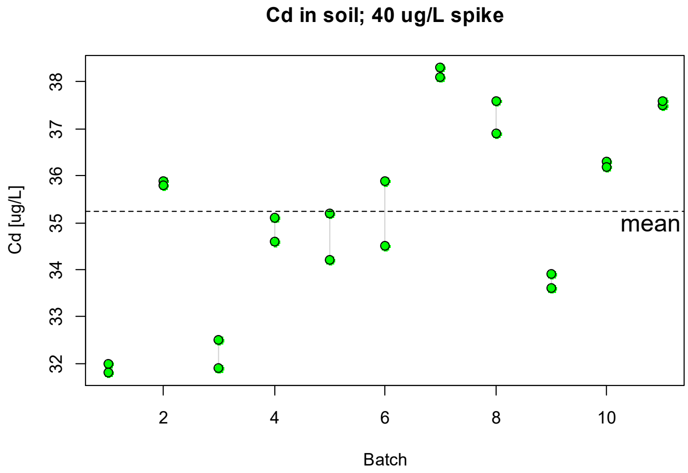
Strategies for reducing validation effort

- Get more than one performance characteristic from a single experiment
- Get more information from one experiment
- Use efficient designs to minimise experiment size

Example 1: Bias from a precision experiment

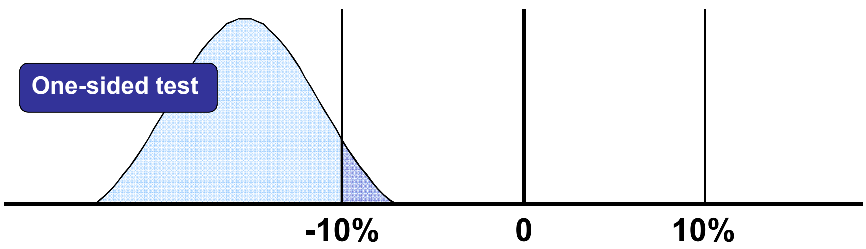
- UK MCERTS soil testing standards set limits for bias (+-10%) and precision (5%) of test methods
- ‘11 x 2’ design recommended
  - 11 days/runs, in duplicate
- 3 soil types, ideally using CRMs
- ANOVA used to determine repeatability and intermediate precision
- Bias checked using a modified *t* test

Example 1 cont.



Example 1 interpretation

- Initial inspection
  - Mean Cd: 35.24 (more than 10% bias)
- Significance test: is bias *significantly* greater than 10%?



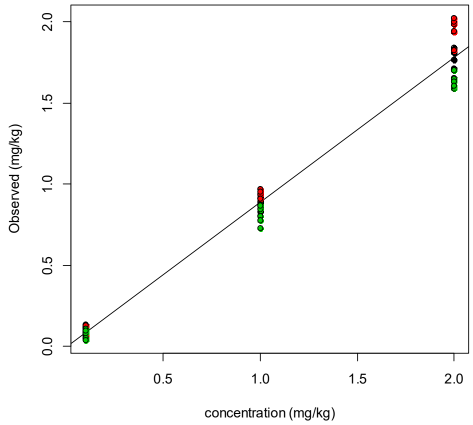
P-value (one-sided, 11* df): 0.31

** Welch-Satterthwaite calculation on ANOVA MS*

Example 2: SANCO precision and detection capability

- 3 runs of 7 observations
- 3 concentrations

- Precision at 3 levels
- Bias at 3 levels
- Linearity review
- Detection capability using ISO 11843



Efficient experiments

Experimental design

Efficient ruggedness designs

- Ruggedness typically requires examination of multiple effects
- Single-effect study needs *n* observations at at least 2 levels
  - 6 effects -> 12n observations
- Factorial designs can be better for small studies
  - But $2^6 = 64$ …

AOAC recommended ruggedness design

Experimental parameter	Experiment number							
	1	2	3	4	5	6	7	8
A or a	A	A	A	A	a	a	a	a
B or b	B	B	b	b	B	B	b	b
C or c	C	c	C	c	C	c	C	c
D or d	D	D	d	d	d	d	D	D
E or e	E	e	E	e	e	E	e	E
F or f	F	f	f	F	F	f	f	F
G or g	G	g	g	G	g	G	G	g
Observed result	s	t	u	v	w	x	y	z

Up to 7 effects in 8 runs

Equivalent to *n* = 4 for 7 parameters

Example problem

- HPLC analysis of Tartrate for monitoring
- Method based on aqueous extraction, SPE cleanup and HPLC
- Factors of interest:
  - Sample size
  - SPE flow rate
  - Additional SPE cleanup stage (is it useful?)
  - LC flow rate
  - LC Column temperature
  - LC Buffer pH

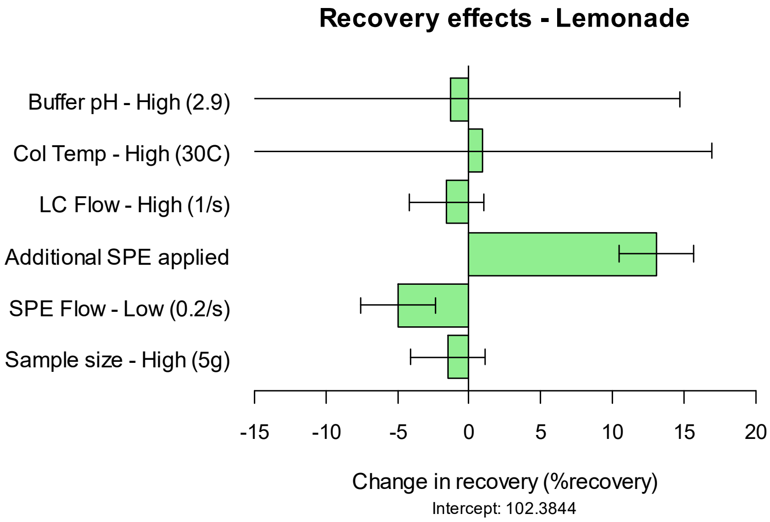
Practical problems

- The basic AOAC design leaves no degrees of freedom
  - and the tartrate design only one
- LC Temperature and buffer pH cannot be changed randomly during a run
  - These four combinations must be in different runs
- "Quick" answer:
  - Four runs allows replication of SPE experiments and leaves a degree of freedom for the LC factors after allowing for run effects

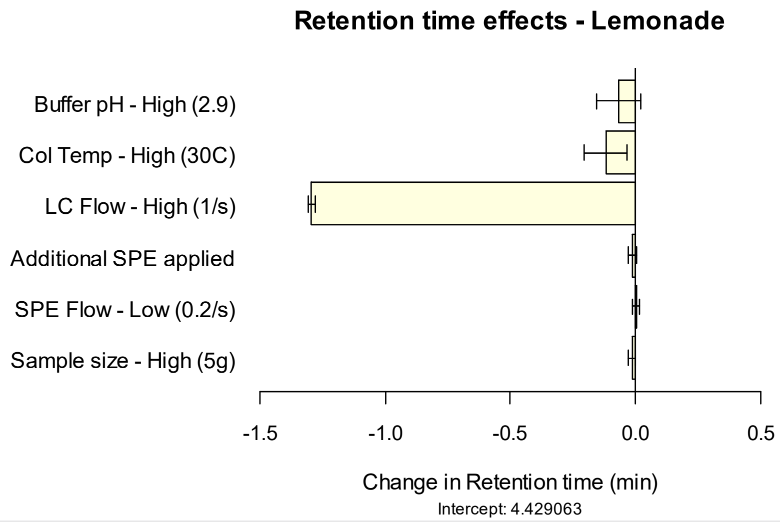
Responses

- Primary interest: Measured tartrate or tartrate recovery
- Also of interest:
  - Is the chromatography likely to be stable?
  - Can we ‘measure’ chromatographic quality at the same time as tartrate?
- Solution: Monitor LC retention time and LC resolution (theoretical plate count) in the same experiment
  - We get the information essentially free

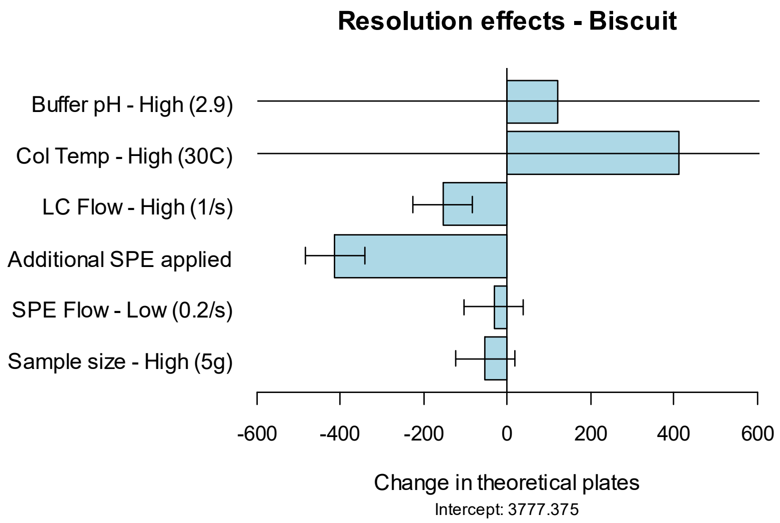
Results - Recovery



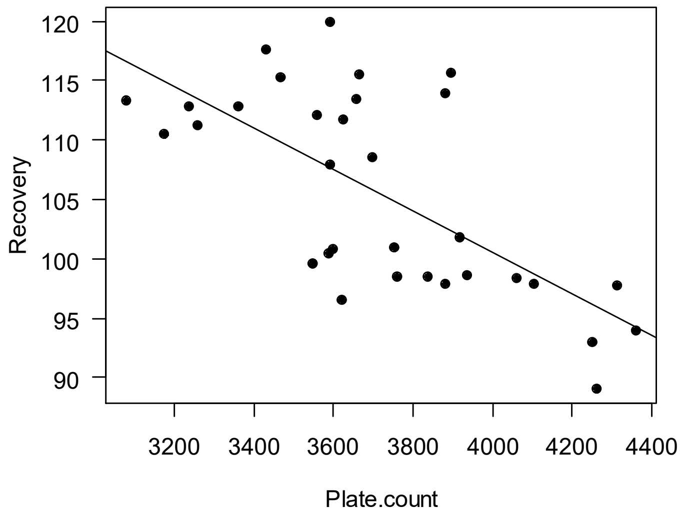
Results – Retention time



Results – LC Resolution



An unexpected bonus



Implications for the tartrate method

- DO use additional SPE cleanup
- DON'T increase the sample size greatly past 2g
- DO keep the LC flow low to keep resolution high
- DO consider checking LC resolution on each sample
  - if the resolution slips, the result may slip with it

Ruggedness test conclusions

- Ruggedness testing isn’t as simple as AOAC make it look
- Monitoring more than one ‘response’ is often simple
- ... and can add a lot to the information available

Conclusions

- Extent of validation is still not harmonised across sectors
- Different guidance still leaves some experiment sizes unclear
- Draft IUPAC Guidance may assist – watch that space!
- Power calculations can help, especially in comparing experiments
- It is possible to ‘work smarter’ for method validation
  - More characteristics per experiment
  - Careful design
  - More ‘responses’ studied